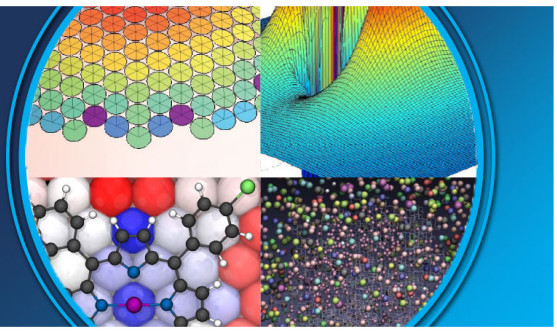
# Accuracy of a Markov state model generated by searching for basin escape pathways

Vijesh J. Bhute, and Abhijit Chatterjee

Citation: The Journal of Chemical Physics **138**, 084103 (2013); doi: 10.1063/1.4792439
View online: https://doi.org/10.1063/1.4792439
View Table of Contents: http://aip.scitation.org/toc/jcp/138/8
Published by the American Institute of Physics





# Accuracy of a Markov state model generated by searching for basin escape pathways

Vijesh J. Bhute[1] and Abhijit Chatterjee[1,2,a)]
[1]*Department of Chemical Engineering, Indian Institute of Technology Kanpur, Kanpur, Uttar Pradesh 208016, India*
[2]*Department of Chemical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India*

(Received 4 July 2012; accepted 4 February 2013; published online 22 February 2013)

Markov state models (MSMs) are employed extensively in literature with the kinetic Monte Carlo (KMC) method for studying state-to-state dynamics in a wide range of material systems. A MSM contains a list of atomic processes and their rate constants for different states of the system. In many situations, only few of the possible atomic processes are included in the MSM. The use of an incomplete MSM with the KMC method can lead to an error in the dynamics. In this work, we develop an error measure to assess the accuracy of a MSM generated using dynamical basin escape pathway searches. We show that the error associated with an incomplete MSM depends on the rate constants missing from the MSM. A procedure to estimate the missing rate constants is developed. We demonstrate our approach using some examples. © *2013 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4792439]

## I. INTRODUCTION

Although the molecular dynamics (MD)[1] method has become an important materials simulation tool for providing insights into the dynamics at short length and time scales, it is computationally prohibitive when rare events involving transitions from one energy basin (or state) to another basin in the potential energy surface (PES) need to be studied. An alternate approach involves the use of a Markov state model (MSM)[2,3] with the kinetic Monte Carlo method (KMC)[4–8] method. The MSM is essentially a "kinetic-map" of the PES describing escape pathways from one basin to another while providing the rate constant associated with the escape. Figure 1 provides a schematic of a 2D PES with several basins and some of the escape pathways. A MSM can be used with KMC to quickly reach length and time scales larger than those accessible to MD by randomly sampling basin escape pathways (or atomic processes) from a catalog of processes, and finding the times associated with each escape. The dynamics from MD and KMC are governed by the same master equation as long as all atomic processes are present in the MSM and they are independent Poisson processes.[9] The KMC dynamics can be incorrect when the MSM is missing processes. This leads to the question: what is the error associated with a MSM model with missing processes when compared to molecular dynamics?

The use of MSMs with KMC is widespread in catalysis, surface science, and biology, however, importance of this question becomes evident once it is realized that finding all processes from a basin is a challenging task. For instance, in most material systems it is usually difficult to guess processes involving concerted movement of more than one atom. Examples of such processes are abound in surface diffusion.[10–21] A more recent approach to overcome guesswork entails constructing the MSM by searching for processes from various basins by performing basin escape pathway search (BEPS).[22–26] Examples of dynamical BEPS approaches include MD,[27–30] accelerated MD,[31–35] and Monte Carlo[36–38] where the processes are found by following the true dynamics of the system. Examples of static BEPS techniques include nudged elastic band[39] and mode sampling[23,40] methods which involve a study of the PES without performing any dynamical calculations. An advantage of BEPS is that the escape pathways from a particular basin can be obtained without guessing the types of processes. Recently, several successful applications of such an approach have been demonstrated in Refs. 3, 27, 28, 41, and 42 where long MD trajectories were employed to build a MSM for biomolecules. However, it important to realize that the MSM constructed using BEPS can still be missing atomic processes. In many situations, it is not evident whether all pathways that are relevant to the dynamics are present in the MSM. Clearly, there is a need to develop a mathematical framework to quantify the error in a MSM.

In this work, we develop an error measure for a MSM generated using dynamical BEPS techniques. The main idea of our approach is that a "detailed" MSM is generated separately using dynamical BEPS techniques. Later, this MSM is supplied to a KMC code to perform KMC dynamics. As we begin performing longer KMC simulations with the fixed MSM, it is likely that one of the processes that is missing in the MSM will have a higher chance of being observed in the correct dynamics. However, the process will never be observed in the KMC dynamics as it is missing from the MSM. We determine the error in the MSM in terms of the probability that processes that are missing in the MSM would have been selected in the correct dynamics. The error in MSM is large when there is the probability is high. Using this

a)Author to whom correspondence should be addressed. Electronic mail: achatter@iitk.ac.in.

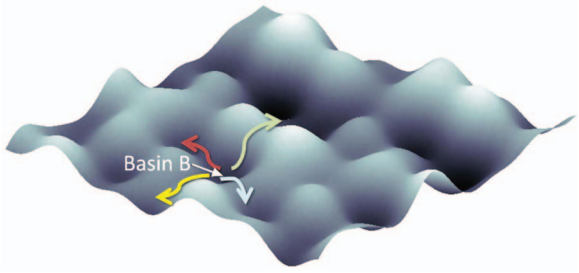



FIG. 1. Schematic of a 2D potential energy surface. Arrows denote escape pathways from a particular basin to other basins in the potential energy surface (PES). The escape pathways correspond to atomic processes in a material system. A Markov state model (MSM) is exact when it contains all basin escape pathways in the PES and their rates.

rationale, we can estimate the duration for which the MSM can be employed with KMC so that the associated error in the KMC dynamics is low. Alternatively, one can determine the error in the KMC dynamics when the MSM is used for certain duration of time.

The paper is divided into the following sections. In Sec. II, we present the rationale behind the error associated with a catalog of processes. The time for which the catalog can be used with KMC, such that the error is less than a prescribed value, is derived. It is shown that the time depends on the rate constants of processes that are missing from the catalog. In Secs. III and IV, a self-consistent procedure for generating a catalog of escapes is developed when none of the escape pathways are known to us initially. In Secs. V and VI, we assess the performance of our procedure by studying three test basins. Finally, conclusions are presented in Sec. VII.

## II. RATIONALE BEHIND THE ERROR MEASURE

In this section, we describe a procedure for building a catalog of processes for a particular basin B in the potential energy surface using dynamic BEPS techniques. Initially, none of the processes from B are known. Basin escapes are sought by performing multiple BEPS calculations, such that each calculation begins in B and the calculation is stopped once the system escapes the basin. A MSM is obtained by repeating this procedure to obtain process catalogs for different basins in the PES.

Although this approach of finding basin escapes is inspired by the temperature accelerated dynamics (TAD) techniques pioneered by Voter,[34,43] there is an important difference between TAD and our procedure. The main goal in TAD is to find an atomic process from the basin with the shortest escape time in a computationally efficient manner. In order to achieve this goal, a MD based BEPS is performed at a temperature that is higher than the system temperature. This enables rare events to occur more frequently than at the original temperature, thus making TAD computationally efficient. Since old escape times cannot be reused, separate BEPS calculations are needed to find the next escape times from the basin. Using the TAD procedure, one can obtain a sequence of basin escapes at the original system temperature (that is correct with a chosen confidence level) and a catalog of processes from the basin. Once this sequence of escapes is available, it is straightforward to study state-to-state transitions just as in standard KMC. However, the TAD implementation is more complicated than the standard KMC approach and requires significantly more computational resources than KMC. One feature that makes KMC attractive is the ease with which millions to billions of escape times can be generated once a catalog of processes from the basin is available. Indeed, the sequence of escape times from different basins obtained using TAD can be provided to KMC, however, such a database prepared for a large number of basins can become unwieldy given that a large fraction of escape times stored may never be used. We can alleviate this problem by supplying only the catalog of processes found using dynamical BEPS to KMC so that new sequence of escapes times can be generated with standard KMC. This forms the basis for our approach. In this work, we perform BEPS to obtain a sequence of escapes at the system temperature, so that the total time elapsed in the basin B while performing BEPS is $t_B$. Note that our approach is general since one can use MD and other dynamical techniques including accelerated MD to obtain an accurate sequence of escape times. We answer the question: what is the error in the KMC dynamics due to the missing rate constants when the catalog is used with KMC to reach a certain KMC time. Alternatively, we can find how long the catalog can be employed with KMC such that the error in the KMC dynamics is less than a maximum error. As it will be clear later, a catalog that has been that generated from a sequence of basin escapes will be accurate for a time that can be much shorter than $t_B$. This aspect raises concerns about the accuracy of most MSMs that are employed with KMC for as long as required. Our approach, which retains the simplicity of the standard KMC by generating escape times from the MSM, can fill this gap by providing the accuracy associated with the dynamics.

The catalog of *known processes* from basin B obtained using BEPS is denoted $C_K$. More processes are added to $C_K$ as the time $t_B$ increases. The catalog of *missing* or *unknown processes* is denoted $C_U$. The complete catalog of processes from B is given by s. Next, we obtain the probability of observing a process in the correct dynamics, i.e., when $C_C$ is used with KMC. In the correct dynamics, the time $\tau$ associated the first escape involving a missing process is exponentially distributed, i.e.,

$$p(\tau) = k_U \exp(-k_U \tau). \tag{1}$$

Here, $k_U$ denotes the sum of rate constants of unknown processes. Using Eq. (1), the probability $p_U$ that at least one process from $C_U$ will be selected during time $\tau$ is given by

$$p_U = \int_0^\tau k_U \exp(-k_U \tau') d\tau' = 1 - \exp(-k_U \tau). \tag{2}$$

A large (small) value of $p_U$ indicates a high (low) probability that one of the missing processes will be selected in the correct dynamics. As long as $p_U$ remains small, the catalog $C_K$ is deemed to contain all processes *relevant* to the dynamics because it is unlikely that a process from $C_U$ will be selected. Hence, $p_U$ is the error measure for the catalog $C_K$.

Alternatively, we can ensure that the probability $p_U$ is less than a maximum error $\delta$. Equation (2) introduces a timescale

$\tau_V$, which we term as the validity time for $C_K$, for which the catalog $C_K$ remains accurate with the error bound given by $\delta$. From Eq. (2), the validity time for $C_K$ is given by

$$\tau_V = -\frac{\ln(1-\delta)}{k_U}. \quad (3)$$

Equations (2) and (3) are the main results of this work. A nice feature of our approach is the validity time depends on the maximum error in KMC dynamics. Once a catalog $C_K$ for basin B has been prepared, it can be reused multiple times with KMC when the system visits B, as long as the KMC time spent in the basin is less than $\tau_V$. When $\tau$ exceeds $\tau_V$, the probability $p_U$ that a missing process will be observed in the correct dynamics is greater than $\delta$. The validity time for the catalog is small when large rate constants are missing in $C_K$, i.e., $k_U$ is large. Equation (3) shows that the validity time also depends on the residence time $t_B$ since it determines the value of $k_U$. The catalog validity time can be increased by performing additional BEPS calculations so that the basin residence time $t_B$ increases. This will result in a catalog that is generated *on-the-fly* with an error bound $\delta$. It can be shown that the maximum error in the KMC dynamics using a MSM containing such catalogs is also $\delta$.

Unfortunately, the sum of rate constants $k_U$ for the unknown processes is not known to us *a priori*. In the remainder of the paper, we develop a procedure for estimating $k_U$ once a catalog $C_K$ is generated for a basin using dynamical BEPS calculations. We rewrite Eq. (3) in terms of the sum of rate constants for the complete and known catalogs, i.e.,

$$\tau_v = -\frac{\ln(1-\delta)}{k_C - k_K}. \quad (4)$$

Maximum likelihood estimation of the sum of rate constants $k_C$ associated with exponentially distributed escape times provides an estimate for $k_C$ given by[3]

$$\tilde{k}_C = \frac{m}{t_B}. \quad (5)$$

Here, m denotes the number of escapes from B. Note that the time $t_B$ is specific to the BEPS calculations used to generate the catalog $C_K$, while $\tau$ is the time elapsed while using the catalog $C_K$ with KMC. Using Eqs. (4) and (5), we obtain an estimate for the validity time

$$\tilde{\tau}_V = -\frac{\ln(1-\delta)}{\tilde{k}_C - k_K}. \quad (6)$$

Here, $k_K$ denotes the sum of rate constants of known processes. We performed dynamic calculations to assess the accuracy of Eq. (5). Instead of performing BEPS for a real material system, standard KMC method was employed with a catalog $C_C$ that is already known to us to seek escape pathways. In reality, when a new material system is being studied the catalog $C_C$ will not be known to us. However, we have used this approach since KMC simulations are orders of magnitude faster than most BEPS techniques and this enables us to assess our procedure for obtaining the catalog validity time. The implementation of our procedure with MD and other dynamical methods will be the subject of a future paper. In the remainder part of this work, we will use KMC-based sampling of a catalog $C_C$ as replacement for BEPS.

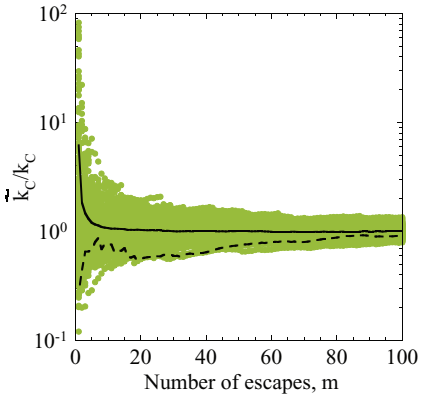



FIG. 2. The sum of rate constants $\tilde{k}_C$ estimated using Eq. (5) when $\theta_{cut}$ is infinite. Filled green circles denote results from 200 catalog generation attempts. The solid line shows the average $\tilde{k}_C$ from the 200 attempts, while the dashed line shows $\tilde{k}_C$ for a particular attempt. It is observed that $\tilde{k}_C > k_C$ in some cases, and $\tilde{k}_C < k_C$ in other catalog generation attempts.

The filled circles in Fig. 2 shows the variation in $\tilde{k}_C$ computed using Eq. (5) for 200 catalog generation attempts, i.e., 200 catalogs for the same basin B is generated using different starting random seeds for our KMC-based BEPS. The number of escapes m and the corresponding residence time $t_B$ is monitored. The time $t_B$ in the BEPS calculations is advanced by $-\ln\xi/k_C$ after each escape from the basin. Here, $\xi$ is a uniform random deviate. When a new process is found using BEPS, it is added to $C_K$. The filled circles show that a large variation in $\tilde{k}_C$ from one catalog to another is present. The values for $\tilde{k}_C$ obtained using Eq. (5) for one catalog generation attempt is shown in Fig. 2 by the dashed line. The solid line in Fig. 2 denotes the average from the 200 attempts. Using Eq. (5), the ratio $\tilde{k}_C/k_C$ is a dimensionless random variable given by

$$\frac{\tilde{k}_C}{k_C} = \frac{m}{\theta_B}, \quad (7)$$

where $\theta_B = k_C t_B = \sum_{j=1}^{m} \ln\xi_j$ is the dimensionless time and j denotes the jth escape. Equation (7) shows that $\tilde{k}_C/k_C$ will depend only on the random seed employed and is independent of the catalog $C_C$ used. In other words, Fig. 2 shows the typical behavior observed with BEPS. We find that $\tilde{k}_C$ slowly converges to the correct value of the rate constant $k_C$ as m increases for all 200 catalogs that were generated. When $\tilde{k}_C > k_C$, a smaller value of the validity time $\tilde{\tau}_V$ is obtained using Eq. (6). Such situations lead to an increase in the computational cost of the method since additional BEPS calculations will be required to reach a target validity time. When $\tilde{k}_C < k_C$, the validity time is larger than what is should be. This will result in an error that is greater than $\delta$. It was observed in some catalog generation attempts in Fig. 2 that the validity time can become negative when $\tilde{k}_C < k_K$ in Eq. (6). Because of these reasons, $\tilde{k}_C < k_C$ is not acceptable to us. In Sec. III, we develop a procedure that ensures that $\tilde{k}_C$ is rarely smaller than $k_C$. This will enable us to employ $\tilde{k}_C$ with Eq. (6) to obtain the validity time for $C_K$.

## III. INTRODUCING A CUT-OFF TIME IN DYNAMICAL BEPS

The value of $\tilde{k}_C$ will be greater than $k_C$ when $m/k_C t_B > 1$ (see Eq. (7)). This is possible when large escape times, which lead to greater increase in $t_B$, are avoided. In order, to achieve this, we introduce the concept of a cut-off time $t_{cut}$ in dynamical BEPS. Any escape that occurs before $t_{cut}$ (the value of $t_{cut}$ will be found later) it is regarded as a successful transition, and will contribute to an increase in the value of m and $t_B$. Escapes that will occur after time $t_{cut}$ are not used for estimating $\tilde{k}_C$. The use of $t_{cut}$ has another advantage when BEPS calculations are performed in parallel. Every once a while, a BEPS calculation can yield a large escape time. Such calculations lower the parallel efficiency of the procedure when an expensive BEPS technique, such as MD, is being used. The parallel efficiency of the procedure becomes higher by preventing BEPS calculations to proceed beyond $t_{cut}$.

We derive the probability density $p(t_B|m,t_{cut})$ associated with $t_B$ given that m successful escapes have occurred while using the cut-off time $t_{cut}$. The probability density for $m = 1$ is given by

$$p(t_B|1, t_{cut}) = \frac{k_C \exp(-k_C t_B)}{\Xi}\{H(t_B) - H(t_B - t_{cut})\}. \quad (8)$$

Here, $\Xi = 1 - \exp(-k_C t_{cut})$ and H(t) is the Heaviside step function. The term in the curly brackets in Eq. (8) ensures that $t_B \in [0, t_{cut})$ for $m = 1$. The probability density that the mth escape occurs at time $t_B$ is given by

$$p(t_B|m, t_{cut}) = \frac{{k_C}^m \exp(-k_C t_B)}{\Xi^m (m-1)!} I_m, \quad (9)$$

where

$$I_m(t|t_{cut}) = \sum_{i=0}^{m} (-1)^i \frac{m!}{i!(m-i)!}(t - i t_{cut})^{m-1} H(t - i t_{cut}). \quad (10)$$

More details are provided in the Appendix. Using a dimensionless time $\theta_{cut} = k_C t_{cut}$, we can rewrite Eq. (9) as

$$p_B(\theta_B|m, \theta_{cut}) = \frac{\exp(-\theta_B)}{\{1 - \exp(-\theta_B)\}^m (m-1)!} \times \sum_{i=0}^{m} (-1)^i \frac{m!}{i!(m-i)!}(\theta_B - i\theta_{cut})^{m-1} H(\theta_B - i\theta_{cut}). \quad (11)$$

When $\theta_{cut} \to \infty$ the probability density for time $t_B$ associated with m escapes is recovered, i.e.,

$$p_B(\theta_B|m) = \frac{{\theta_B}^{m-1} \exp(-\theta_B)}{(m-1)!}. \quad (12)$$

Figure 2 corresponds to the case where $\theta_{cut}$ is infinite.

Figure 3 shows the effect of introducing a shorter cut-off time using $\theta_{cut} = 1$ while performing BEPS using the same random number seeds used in Fig. 2. The cut-off time $t_{cut}$ can be obtained since we already know $k_C$ for our catalog. More discussion on finding $k_C$ and $\tilde{k}_C$, which are typically unknown *a priori*, will be provided in Sec. IV. It is observed that $\tilde{k}_C > k_C$ for all 200 catalog generation attempts. As in the case of Fig. 2, any complete catalog will exhibit similar behavior due to the use of dimensionless times. The solid line shows the average value of $\tilde{k}_C$ (averaged over the 200 attempts) has a systematic deviation and it will never converge to $k_C$. The dashed line shows the value of $\tilde{k}_C$ for a particular attempt.

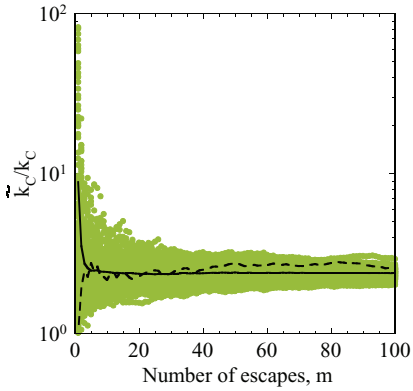



FIG. 3. The sum of rate constants $\tilde{k}_C$ estimated using Eq. (5) when the cut-off time $\theta_{cut} = 1$. The green filled circles denote results from 200 catalog generation attempts. Unlike Fig. 2, $\tilde{k}_C$ is always greater than $k_C$. The solid line shows the average $\tilde{k}_C$ from the 200 attempts, while the dashed line shows $\tilde{k}_C$ for a particular attempt.

In order to explain this systematic deviation in $\tilde{k}_C$, we analyze the probability density for escape times for $m = 10$ escapes in Fig. 4 using $\theta_{cut} \to \infty$ and $\theta_{cut} = 1$. The histogram from BEPS calculations and the dashed lines denote the probability density from Eq. (11) are in excellent agreement. In Fig. 4(a), it is observed that the probability of $\theta_B > m$ (region in the shaded part) is large when $\theta_{cut} \to \infty$. This explains why in some catalog generation attempts we find that $\tilde{k}_C < k_C$ when $\theta_{cut} \to \infty$. When $\theta_{cut} = 1$, the probability density shifts to the left. We find that the probability density remains to the left of $\theta_B = m$ as m increases. This behavior explains the

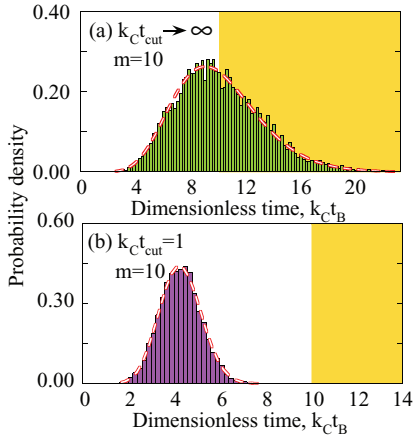



FIG. 4. Results from BEPS calculations (histogram) and Eq. (11) (dashed lines) are in good agreement as shown for $m = 10$ escapes from the basin using dimensionless cut-off times $\theta_{cut}$: (a) infinite, (b) 1. Here, $\theta_{cut} = k_C t_{cut}$, $k_C$ is the sum of rate constants for the complete catalog of processes from B. The shaded region (in orange) indicates situations that lead to $\tilde{k}_C < k_C$.

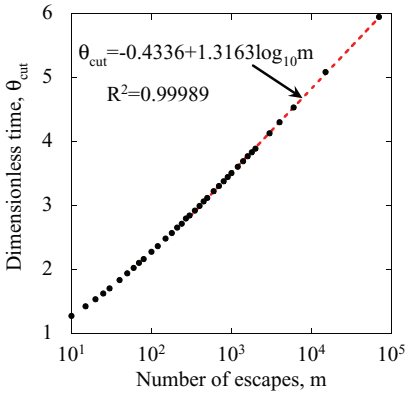



FIG. 5. The value of the dimensionless time $\theta_{cut}$ in Eq. (13) that results in probability $P_e = 10^{-5} \pm 10^{-6}$. Here, m is the number of escapes from basin B. Dashed red line denotes fit obtained for larger values of m.

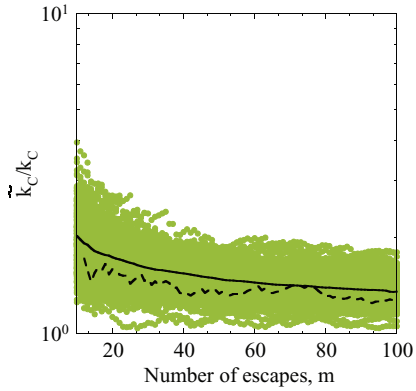



FIG. 6. The value of $\tilde{k}_C$ appears to converge to $k_C$ when the cut-off time $\theta_{cut}$ from Fig. 5 is employed. It is also found that $\tilde{k}_C$ is always greater than $k_C$ as shown by the green filled circles that are results obtained from 200 catalog generation attempts. Solid line denotes the average value of $\tilde{k}_C$ (averaged over the 200 attempts) while the dashed line shows $\tilde{k}_C$ for a particular attempt.

systematic deviation in the value of $\tilde{k}_C$ as witnessed in Fig. 3. Next, we obtain the value of $t_{cut}$ which enables safe estimation of $\tilde{k}_C$ and correct convergence to $k_C$.

### A. Numerical evaluation of $\theta_{cut}$

We find a value of $\theta_{cut}$ that ensures that $\theta_B > m$ occurs with a low probability $P_e$, i.e.,

$$\int_m^\infty p_B(\theta_B|m, \theta_{cut})d\theta_B = P_e. \quad (13)$$

The cut-off time $\theta_{cut}$ was obtained from Eq. (13) using BEPS calculations for different values of m such that the resulting $P_e = 10^{-5} \pm 10^{-6}$. Figure 5 shows $\theta_{cut}$ for number of escapes reaching 70 000. It is found that the cut-off time depends on the number of escapes m. A logarithmic fit is obtained for the cut-off time in Fig. 5 for $m > 300$ with a Pearson correlation coefficient $R^2 = 0.99989$. Other values of $\theta_{cut}$ for $m < 300$ are given in Table I.

TABLE I. Cut-off time $\theta_{cut}$ (for different number of escapes m) that ensures $P_e = 10^{-5} \pm 10^{-6}$ in Eq. (13).

| Number of escapes m | Cut-off time $\theta_{cut}$ |
|---|---|
| 10 | 1.28 |
| 15 | 1.43 |
| 20 | 1.54 |
| 25 | 1.63 |
| 30 | 1.71 |
| 40 | 1.84 |
| 50 | 1.95 |
| 60 | 2.03 |
| 70 | 2.11 |
| 80 | 2.17 |
| 100 | 2.28 |
| 120 | 2.37 |
| 150 | 2.49 |
| 180 | 2.57 |
| 210 | 2.66 |
| 240 | 2.72 |
| 270 | 2.80 |
| 300 | 2.85 |

Figure 6 shows $\tilde{k}_C/k_C$ using values of $\theta_{cut}$ from Fig. 5 as m increases. It is observed that unlike Fig. 2, $\tilde{k}_C$ is always greater than $k_C$. Furthermore, unlike Fig. 3 the solid line shows the value of $\tilde{k}_C$ averaged over the 200 catalog generation attempts appears to converge to the correct value of $k_C$. The dashed line shows $\tilde{k}_C$ for a particular attempt. However, one obstacle that remains is that $k_C$ used to define the dimensionless times is generally not available to us. In Sec. IV, we develop a self-consistent procedure to obtain $\tilde{k}_C$ and $t_{cut}$.

## IV. A SELF-CONSISTENT PROCEDURE FOR ESTIMATING THE CUT-OFF TIME AND RATE CONSTANTS

The procedure outlined in Sec. III requires the knowledge of the cut-off time $t_{cut}$ for finding $\tilde{k}_C$ using Eq. (5). As we have seen in Sec. III, the value of $t_{cut}$ depends on the sum of rates $\tilde{k}_C$ (assuming $\tilde{k}_C$ is close to $k_C$). However, $\tilde{k}_C$ and $k_C$ are usually not known to us in the beginning. In this section, we develop a procedure to determine $\tilde{k}_C$ and $t_{cut}$ in a self-consistent manner.

The first step involves obtaining an initial estimate for $\tilde{k}_C$. The catalog $C_K$ is created from certain number of initial BEPS calculations and the value of $\tilde{k}_C$ is set to $k_K$. Subsequently, using an iterative procedure the cut-off time $t_{cut}$ is obtained from $\theta_{cut}(m)/\tilde{k}_C$ using Fig. 5 and $\tilde{k}_C$ is estimated from Eq. (5), such that m, $t_B$, $t_{cut}$, and $\tilde{k}_C$ are self-consistent. The validity time for the catalog $C_K$ is obtained using Eq. (6). Additional BEPS calculations are required when the validity time needs to be extended. In our implementation, we continuously update the values of m and $t_B$, for a list of N target number of escapes arranged in an ascending order $\{m_1^T, m_2^T, ..., m_N^T\}$. For instance, in our calculations we set $m_1^T = 50$ and $m_N^T = 70\,000$. Our goal is to find the time $t_B$ once a target number of successful escapes is reached. The cut-off time for the targeted number of escapes is obtained from Fig. 5 and is given by $\{\theta_1, \theta_2, ..., \theta_N\}$. Only escapes that have occurred before the cut-off time $\theta_i/\tilde{k}_C$ are used for obtaining the validity time after $m_i^T$ number of escapes have occurred.

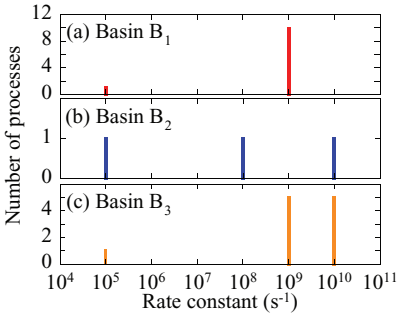



FIG. 7. Histogram showing the rate constants of processes in the complete catalog $C_C$ for three different basins used to assess the performance of our procedure. (a) Basin $B_1$, (b) basin $B_2$, (c) basin $B_3$.

## V. ASSESSMENT OF THE VALIDITY TIME

In this section, we study the validity time for catalogs generated for three test basins. The types of processes in the catalog $C_C$ associated with these basins and their rates as shown in Fig. 7.

### A. Test basin $B_1$

Basin $B_1$ consists of one slow and 10 fast processes, such that there is four orders of magnitude separation in the rates of the slow and fast processes. Figure 7(a) shows the rate constants in the catalog $C_C$. The sum of rate constants of the fast processes is $10^{10}$ $s^{-1}$.

Figure 8(a) shows that the validity time $\tilde{\tau}_V$ (with $\delta = 0.1$) for $C_K$ increases as the residence time $t_B$ increases. At time $t_B = 0$, the validity time $\tilde{\tau}_V$ is set to zero. The first value of validity time obtained is 0.01147 ns after m = 50 (since the smallest target ${m_1}^T$ was set to 50). The value of $\tilde{\tau}_V$ fluctuates due to the randomness present in $t_B$. The new value of $\tilde{\tau}_V$ is prevented from decreasing by comparing it to the current value of $\tilde{\tau}_V$. This explains the step-increase in $\tilde{\tau}_V$ in Fig. 8(a). Figure 8(a) shows that the time scales $\tilde{\tau}_V$ and $t_B$ differ by three orders of magnitude for basin $B_1$. This difference can be qualitatively explained in terms of the estimate for the unknown rate. A simple estimate for $\tilde{k}_U$ can be obtained by considering the scenario where an escape pathway that has not been observed so far, will be observed shortly if additional BEPS is performed. The rate for this pathway will be $O(1/t_B)$, where O( ) denotes order of magnitude. Using $\tilde{k}_U \approx 1/t_B$ in Eq. (6), we find that $\tilde{\tau}_V \approx -t_B \ln(1-\delta)$. When $\delta = 0.1$, we obtain $\tilde{\tau}_V \approx 0.1 t_B$. This shows that the validity time $\tilde{\tau}_V$ scales linearly with the residence time $t_B$ as is evident from Fig. 8(a).

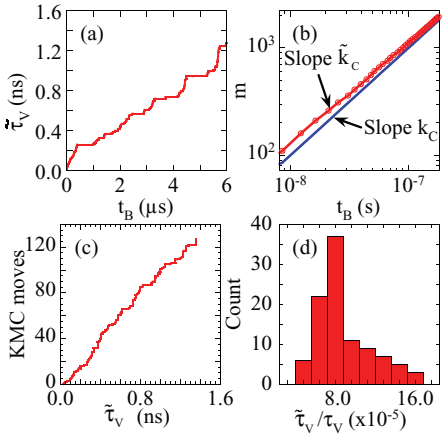



FIG. 8. Results obtained when a catalog $C_K$ is generated on-the-fly for the complete catalog in Fig. 7(a): (a) Catalog validity time $\tilde{\tau}_V$ plotted against the basin residence time $t_B$, (b) convergence of $\tilde{k}_C$ as $t_B$ increases, and (c) number of KMC moves with $\tilde{\tau}_V$. (d) Histogram for the ratio $\tilde{\tau}_V/\tau_V$ for catalogs generated with m = 1000 for 100 catalog generation attempts.

Figure 8(b) shows the value of m and $t_B$ obtained from successful BEPS transitions. It is observed that the red line showing number of escapes m approaches the blue line (with slope $k_C$) confirming that our scheme ensures convergence of $\tilde{k}_C$ to $k_C$ while preventing $m/t_B$ from being smaller than $k_C$. This behavior is similar to the one shown in Fig. 6, however, here the value of m reaches 70 000.

As $C_K$ was being generated on-the-fly, KMC dynamics was followed using the catalog $C_K$, such that the KMC time $\tau$ spent in the basin is less than $\tilde{\tau}_V$. Each time the KMC method selects pathway process from $C_K$, the system is returned to basin $B_1$. In this way, we studied the catalog validity time for a large number of KMC moves. Additional BEPS calculations are performed when the KMC time exceeds the catalog validity time. Figure 8(c) shows that the number of KMC moves increases linearly with the validity time $\tilde{\tau}_V$ as the catalog it is being generated *on-the-fly*.

The efficiency of our procedure depends on the rate of convergence of $\tilde{k}_C$ towards $k_C$. Since the catalog $C_C$ for this basin is already known to us, we can compute the actual validity time for catalog as it is being generated for basin $B_1$. Figure 8(d) shows the histogram for $\tilde{\tau}_V/\tau_V$ obtained from 100 different catalogs generation attempts using our procedure. It is observed that the average value of $\tilde{\tau}_V/\tau_V$ is $8.85 \times 10^{-5}$ (averaged over 100 catalog generation attempts), i.e., the time $\tilde{\tau}_V$ is orders of magnitude smaller than $\tau_V$. The ratio of the estimated and the actual catalog validity times is given by

$$\frac{\tilde{\tau}_V}{\tau_V} = \frac{k_U}{m/t_B - k_K}. \tag{14}$$

In the case of basin $B_1$, the separation of time scales between the fast and slow pathways is $10^5$ times. At m = 100, all fast processes are known and $k_K$ is close to $k_C$, however, most of the slow processes still remain to be observed with BEPS. From Fig. 6, which is valid for any catalog, it is evident that the difference $m/t_B - k_K \approx O(k_C)$. Using $k_C \approx k_K$, we find that $\tilde{\tau}_V/\tau_V \approx k_U/k_K$, i.e., the separation of time scales determines the $\tilde{\tau}_V/\tau_V$. After m = 1000, we find that for basin $B_1$ $k_U$ is the smallest rate constant and $\tilde{k}_U \approx k_K/10$. This explains why the separation in the time scales $\tilde{\tau}_V$ and $\tau_V$ is approximately $10^{-4}$ in Fig. 8(d).

### B. Test basin $B_2$

Basin $B_2$ consists of three types of processes with time scale separation of two orders of magnitude or more. Figure 7(b) shows the rate constants in the catalog $C_C$. Figure 9(a) shows the validity time plotted against the basin residence time $t_B$ for a particular catalog. The random seed used in Fig. 8(a) is also used in Fig. 9(a). Since the sum of rate

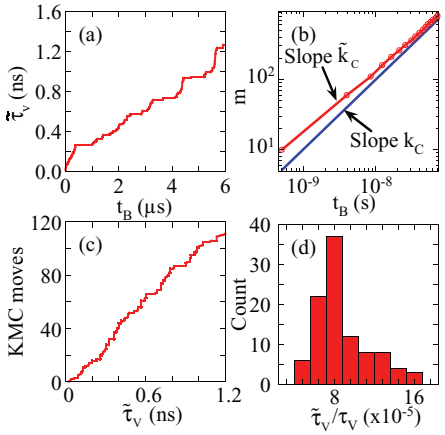



FIG. 9. Results obtained when a catalog $C_K$ is generated on-the-fly for the complete catalog in Fig. 7(b): (a) Catalog validity time $\tilde{\tau}_V$ plotted against the basin residence time $t_B$, (b) convergence of $\tilde{k}_C$ as $t_B$ increases, and (c) number of KMC moves with $\tilde{\tau}_V$. (d) Histogram for the ratio $\tilde{\tau}_V/\tau_V$ for catalogs generated with m = 1000 for 100 catalog generation attempts.

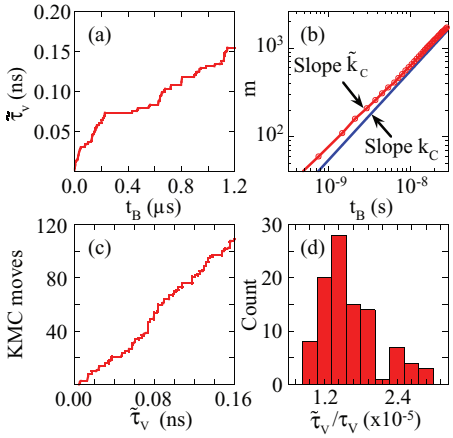



FIG. 10. Results obtained when a catalog $C_K$ is generated on-the-fly for the complete catalog in Fig. 7(c): (a) Catalog validity time $\tilde{\tau}_V$ plotted against the basin residence time $t_B$, (b) convergence of $\tilde{k}_C$ as $t_B$ increases, and (c) number of KMC moves with $\tilde{\tau}_V$. (d) Histogram for the ratio $\tilde{\tau}_V/\tau_V$ for catalogs generated with m = 1000 for 100 catalog generation attempts.

constants for basins $B_1$ and $B_2$ is nearly same Figs. 8(a) and 9(a) appear to be identical. As in the case of $B_1$, the first validity time obtained is 0.011 ns (after m = 50 escapes). The plot for convergence of $\tilde{k}_C$ in Fig. 9(b) and the number of KMC moves in terms of $\tilde{\tau}_V$ in Fig. 9(c) also appear similar to the ones in Fig. 8. The difference in the two basins mainly arises from the presence of a pathway with rate constant in $10^8$ $s^{-1}$ in basin $B_2$. The largest two rates are observed within 1000 KMC moves, as a result, as in the case of Fig. 8(d), we find that for basin $B_2$ $k_U$ is usually the smallest rate constant while $\tilde{k}_U \approx k_K/10$. Figure 9(d) shows that $\tilde{\tau}_V/\tau_V \approx 10^{-4}$ after m = 1000.

### C. Test basin $B_3$

Basin $B_3$ consists of three sets of rate constants as shown in Fig. 7(c). The sum of the rate constants of the fastest processes is $5 \times 10^{10}$ $s^{-1}$. Nine out of 11 processes were added to the catalog $C_K$ after 100 escapes for the catalog generation attempt shown in Fig. 10(a). The value of $k_K$ is very close to that of $k_C$ at this stage. The first validity time obtained for this basin is 2.025 ps for m = 50. Basin escapes were observed earlier (shorter $t_B$) since the sum of all the rates is greater than the ones for basins $B_1$ and $B_2$. Figures 10(b) and 10(c) show the convergence of $\tilde{k}_C$ with m and the number of KMC moves for $\tilde{\tau}_V$, respectively. The behavior is similar to the one observed in Fig. 8. The average value of the ratio $\tilde{\tau}_V/\tau_V$ is 1.62 $\times 10^{-5}$ after m = 1000 (averaged over 100 independent catalog generation runs). As in the case of Fig. 8, this value can be explained in terms of the separation of time scales between $k_U$ and $k_K$.

We find that even though the types of processes and their rates were different for the three basins considered in this section the conclusion remains the same, namely, (i) the validity time $\tilde{\tau}_V$ is always less than $t_B$, and (ii) the validity time $\tilde{\tau}_V$ can be significantly smaller than the true validity time $\tau_V$. The latter conclusion implies that our procedure yields a catalog $C_K$ that has a level of accuracy much higher than the prescribed level.

## VI. ASSESSING THE ERROR MEASURE AND IMPROVING COMPUTATIONAL EFFICIENCY

In this section, we demonstrate that the procedure developed in Sec. IV can correctly generate a catalog with error less than $\delta$. In addition, we exploit the understanding developed in Sec. V regarding $\tilde{\tau}_V$ to improve the computational efficiency of the procedure. Finally, we discuss how our approach can be implemented with different dynamical techniques to generate accurate MSM models.

The accuracy of a catalog $C_K$ that has already been generated is assessed by performing KMC calculations with the catalog $C_C$ till the validity time $\tilde{\tau}_V$ is reached and finding whether any process is selected that is missing from $C_K$. Note that the KMC calculation with $C_C$ gives a correct dynamical trajectory. The catalog $C_K$ is deemed to have failed when a missing pathway is observed in the KMC calculation. As mentioned is Sec. II, the probability that a particular catalog $C_K$ will fail is $\delta$.

We performed KMC simulations of duration $\tilde{\tau}_V$ with a catalog that was generated in Sec. IV and found the percentage of times the catalog fails. The procedure is repeated for 100 different catalogs generation attempts and the maximum percentage of failing $\delta_{obs}$ is shown in Table II. Three values of $\delta$, namely, 1%, 5%, and 10% are considered. The time $\tilde{\tau}_V$ is obtained for different values of m (m = 100, 1000, 10 000). The study is performed for the three basins described in Sec. V. Note that smaller value of $\delta$ implies a higher accuracy level. It is observed that the maximum probability of failing $\delta_{obs}$ is much less than the specified value of $\delta$ (values denoted by $\Omega = 1$ correspond to present case; the meaning of $\Omega$ will become clear later). For example, $\delta_{obs}$ is 0 for basin $B_1$ and 0.035 for basin $B_3$ when m = 100 and $\delta = 0.1$. This behavior can be explained in terms of the small value of $\tilde{\tau}_V/\tau_V$ due the separation of time scales involving rates from

TABLE II. Error associated with catalogs $C_K$ generated for three different basins in Fig. 7. Ω denotes the multiplying factor used to scale up the catalog validity time.

| Basin $B_1$ | | | | | | |
|---|---|---|---|---|---|---|
| m → | 100 | | 1000 | | 10 000 | |
| Ω → | 1 | 10 000 | 1 | 5000 | 1 | 1750 |
| δ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) |
| 1 | 0.0 | 1.3 | 0.0 | 1.1 | 0.0 | 1.0 |
| 5 | 0.0 | 3.7 | 0.0 | 4.5 | 0.0 | 4.2 |
| 10 | 0.0 | 8.8 | 0.1 | 9.3 | 0.0 | 8.1 |
| Basin $B_2$ | | | | | | |
| m → | 100 | | 1000 | | 10 000 | |
| Ω → | 1 | 15 | 1 | 5000 | 1 | 1750 |
| δ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) |
| 1 | 0.1 | 0.8 | 0 | 1.1 | 0 | 1.0 |
| 5 | 0.3 | 4.4 | 0 | 4.6 | 0.1 | 4.2 |
| 10 | 0.6 | 8.6 | 0.1 | 9.3 | 0 | 8.1 |
| Basin $B_3$ | | | | | | |
| m → | 100 | | 1000 | | 10 000 | |
| Ω → | 1 | 3 | 1 | 20000 | 1 | 8000 |
| δ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) | $\delta_{obs}$ (%) |
| 1 | 0.6 (1.8) | 1.0 | 0 (1.8) | 1.2 | 0 (1.8) | 1.0 |
| 5 | 1.8 (6.6) | 4.5 | 0 (6.7) | 3.3 | 0 (6.7) | 3.9 |
| 10 | 3.5 (11.8) | 9.5 | 0.1 (12) | 6.6 | 0 (12) | 7.2 |

the basin as observed in Sec. V. In some situations where we have an estimate of the separation of time scales denoted by Ω, we can multiply $\tilde{\tau}_V$ with Ω to obtain a more realistic value of the validity time. In Table II, we use $\Omega\tilde{\tau}_V$ as the validity time, such that $\Omega > 1$, and find the error $\delta_{obs}$. It is observed that by choosing a value of Ω, which is close to $\tilde{k}_U/k_U$, the maximum error $\delta_{obs}$ remains close to δ. Note that Ω is small for m = 100 in basins $B_2$ and $B_3$ because the separation in rate constants for the known and missing pathways in these basins is small (see discussion in Sec. V). However, this value increases when m = 1000. We observe that Ω decreases for basins $B_1$, $B_2$, and $B_3$ as m increases from 1000 to 10 000. This can be explained by the fact that $\tilde{k}_C$ is estimated with higher accuracy for larger m, since majority of the fast pathways have been observed in these basins and only the slowest pathways remain to be observed. Hence, the ratio $\tilde{k}_U/k_U$ becomes smaller.

In this work, we have employed KMC as a BEPS technique. When one employs MD instead, the computational cost for generating the MSM will increase by orders of magnitude. The MSM can be generated more efficiently by performing parallel, independent MD based basin searches when multiple processors are available. With the availability of 100–1000 s of processors becoming common nowadays, this would reduce the cost of the MD based MSM generation drastically. Alternatively, one can employ more efficient dynamical techniques, such as accelerated MD, to generate a sequence of escapes from the basin more efficiently than possible with the MD method and then use our approach to obtain the accuracy of the resulting catalog. This will also help in finding slower processes that are typically not accessible to MD. The use of MD and other dynamical approaches to generate accurate MSMs will be a subject of study in our future publications.

## VII. CONCLUSIONS

MSM are used extensively with KMC in literature, but rarely is the correctness of a MSM questioned. In this work, we lay the foundations to develop a systematic procedure for building *accurate* MSMs. The procedure involves finding a sequence of escapes times and a catalog of processes from a basin using dynamical BEPS. The MSM can be used directly with KMC to generate multiple dynamical trajectories without requiring any additional BEPS calculations as long as the basins visited in the dynamics have been studied with BEPS. Unfortunately, the MSM can be missing some processes that are relevant to the dynamics. Hence, there is an error in the dynamics when an incomplete catalog is used with a KMC method. The probability of observing a process that is missing from the catalog in the correct dynamics of the system determines the error. Since this probability depends on the time for which the catalog will be used with KMC, we derive an expression for the catalog validity time that ensures that the error will be less than a prescribed value. The catalog validity time depends on the maximum allowable error and the sum of the unknown rates. We provide a procedure to estimate the unknown rates. We show that the use of a cut-off time with the BEPS calculation allows safe estimation of the validity time. Thus, our approach provides a systematic and general way of building MSMs using BEPS techniques with a controlled accuracy starting from the interatomic potential of a material system. We have demonstrated that further increase in computational efficiency can be obtained by exploiting the time scale separation present in the rate constants for a particular basin. In some cases, the validity time is found to increase by orders of magnitude by exploiting such an idea. Finally, we conclude that the practice of using a MSM with KMC for as long as required can result in large errors. We hope that this work provides a starting point to address this major issue.

## ACKNOWLEDGMENTS

We acknowledge helpful discussions with A. F. Voter. A.C. acknowledges support from BRNS Young Scientist Award from Department of Atomic Energy (DAE-BRNS) No. 2011/36/43-BRNS/1975.

## APPENDIX: PROBABILITY DENSITY FOR TIME REQUIRED FOR GIVEN NUMBER OF ESCAPES

The probability density function $p(t_B|m, t_{cut})$, $m > 1$, obeys

$$p(t_B|m, t_{cut}) = \int_0^{t_B} p(t|m, t_{cut})p(t_B - t|1, t_{cut})dt. \quad (A1)$$

Here, $t_{cut}$ is the cut-off time used for all BEPS calculations in the basin. Equation (8) provides the expression for $p(t_B|1, t_{cut})$. The following list of equations are used to derive Eq. (9):

$$\int_0^\tau \sum_{i=0}^{m} a_i t^i H(t-\alpha)H(t-\beta)dt$$

$$= \sum_{i=0}^{m} \frac{a_i}{i+1}(\tau^{i+1} - \alpha^{i+1}H(\alpha-\beta) - \beta^{i+1}H(\beta-\alpha)), \quad \text{(A2)}$$

$$\int_0^\tau \sum_{i=0}^{m} a_i t^i H(t-\alpha)dt = \sum_{i=0}^{m} \frac{a_i}{i+1}(\tau^{i+1} - \alpha^{i+1}), \quad \text{(A3)}$$

$$\int_0^\tau \sum_{i=0}^{m} a_i t^i H(t-\alpha)H(\beta-t)dt$$

$$= \sum_{i=0}^{m} \frac{a_i}{i+1}(\beta^{i+1} - \alpha^{i+1})H(\beta-\alpha), \quad \text{(A4)}$$

$$\int_0^\tau \sum_{i=0}^{m} a_i t^i H(\alpha-t)H(\beta-t)dt$$

$$= \sum_{i=0}^{m} \frac{a_i}{i+1}(\alpha^{i+1}H(\beta-\alpha) - \beta^{i+1}H(\alpha-\beta)), \quad \text{(A5)}$$

$$\int_0^\tau \sum_{i=0}^{m} a_i t^i \{H(t-\alpha) - H(t-\beta)\}dt$$

$$= \sum_{i=0}^{m} \frac{a_i}{i+1}(\tau^{i+1} - \alpha^{i+1} - \beta^{i+1}), \quad \text{(A6)}$$

$$\int_0^\tau t^i H(t-\gamma)\{H(\alpha-t) - H(t-\beta)\}dt$$

$$= \frac{1}{i+1}\{(\alpha^{i+1} - \gamma^{i+1})H(\alpha-\gamma) + \gamma^{i+1}H(\gamma-\beta)$$
$$+ \beta^{i+1}H(\beta-\gamma) - \tau^{i+1}\}, \quad \text{(A7)}$$

$$\int_0^\tau t^i H(t-\gamma)\{H(\alpha-t) - H(\beta-t)\}dt$$

$$= \frac{1}{i+1}\{(\alpha^{i+1} - \gamma^{i+1})H(\alpha-\gamma)$$
$$- (\beta^{i+1} - \gamma^{i+1})H(\beta-\gamma)\}, \text{ and} \quad \text{(A8)}$$

$$\int_0^\tau (t-\gamma)^i H(t-\gamma)\{H(\alpha-t) - H(\beta-t)\}dt$$

$$= \frac{1}{i+1}\{\gamma^{i+1}(H(\beta-\gamma) - H(\alpha-\gamma)) + \alpha^{i+1}H(\alpha-\gamma)$$
$$- \beta^{i+1}H(\beta-\gamma)\}. \quad \text{(A9)}$$


[1] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford Science, Oxford, 1989).
[2] D. J. Wales, Mol. Phys. **100**(2), 3285 (2002).
[3] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, J. Chem. Phys. **131**, 124101 (2009).
[4] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz, J. Comput. Phys. **17**, 10 (1975).
[5] D. T. Gillespie, J. Comput. Phys. **22**, 403 (1976).
[6] A. F. Voter, in *Radiation Effects in Solids*, edited by K. E. Sickafus, E. A. Kotomin, and B. P. Uberuaga (Springer, Dordrecht, 2006).
[7] A. Chatterjee and D. G. Vlachos, J. Comput.-Aided Mater. Des. **14**(2), 253 (2007).
[8] K. A. Fichthorn and W. H. Weinberg, J. Chem. Phys. **95**, 1090 (1991).
[9] N. G. V. Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, 1992).
[10] P. J. Feibelman, Phys. Rev. Lett. **65**(6), 729–732 (1990).
[11] C. Chen and T. T. Tsong, Phys. Rev. Lett. **64**, 3147 (1990).
[12] P. Stoltze and J. K. Norskov, Phys. Rev. B **48**, 5607 (1993).
[13] M. Villarba and H. Jónsson, Phys. Rev. B **49**, 2208 (1994).
[14] J. C. Hamilton, M. S. Daw, and S. M. Foiles, Phys. Rev. Lett. **74**, 2760 (1995).
[15] O. M. Braun and R. Ferrando, Phys. Rev. E **65**, 061107 (2002).
[16] F. Montalenti and R. Ferrando, Phys. Rev. Lett. **82**, 1498 (1999).
[17] T. R. Linderoth, S. Horch, L. Petersen, S. Helveg, E. Lægsgaard, I. Stensgaard, and F. Besenbacher, Phys. Rev. Lett. **82**, 1494 (1999).
[18] G. Antczak and G. Ehrlich, Phys. Rev. B **71**, 115422 (2005).
[19] G. Antczak, Phys. Rev. B **74**, 153406 (2006).
[20] J. Ferrón, R. Miranda, and J. J. d. Miguel, Phys. Rev. B **79**, 245407 (2009).
[21] Y. Tiwary and K. A. Fichthorn, Phys. Rev. B **81**, 195421 (2010).
[22] D. Konwar, V. J. Bhute, and A. Chatterjee, J. Chem. Phys. **135**, 174103 (2011).
[23] G. Henkelman and H. Jonsson, J. Chem. Phys. **115**, 9657 (2001).
[24] A. Kara, O. Trushin, H. Yildirim, and T. S. Rahman, J. Phys.: Condens. Matter **21**, 084213 (2009).
[25] L. Xu and G. Henkelman, J. Chem. Phys. **129**, 114104 (2008).
[26] L. K. Béland, P. Brommer, F. El-Mellouhi, J.-F. Joly, and N. Mousseau, Phys. Rev. E **84**(4), 046704 (2011).
[27] G. R. Bowman, X. Huang, and V. S. Pande, Cell Res. **20**, 622 (2010).
[28] G. R. Bowman and V. S. Pande, Proc. Natl. Acad. Sci. U.S.A. **107**(24), 10890 (2010).
[29] G. H. Gilmer, H. C. Huang, T. D. de la Rubia, J. Dalla Torre, and F. Baumann, Thin Solid Films **365**(2), 189 (2000).
[30] R. Elber, Curr. Opin. Struct. Biol. **15**(2), 151 (2005).
[31] A. F. Voter, Phys. Rev. B **57**, R13985–R13988 (1998).
[32] A. F. Voter, J. Chem. Phys. **106**(11), 4665 (1997).
[33] R. Miron and K. A. Fichthorn, J. Chem. Phys. **119**(12), 6210 (2003).
[34] M. R. Sorenson and A. F. Voter, J. Chem. Phys. **112**(21), 9599 (2000).
[35] R. A. Miron and K. A. Fichthorn, Phys. Rev. Lett. **93**(12), 128301 (2004).
[36] L. R. Pratt, J. Chem. Phys. **85**(9), 5045 (1986).
[37] C. Dellago, P. G. Bolhuis, and D. Chandler, J. Chem. Phys. **108**(22), 9236 (1998).
[38] D. M. Zuckerman and T. B. Woolf, J. Chem. Phys. **111**(21), 9475 (1999).
[39] G. Henkelman, B. P. Uberuaga, and H. Jónsson, J. Chem. Phys. **113**(22), 9901 (2000).
[40] N. Mousseau and G. T. Barkema, Phys. Rev. E **57**, 2419 (1998).
[41] N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[42] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys. **134**(17), 174105 (2011).
[43] F. Montalenti and A. F. Voter, J. Chem. Phys. **116**, 4819 (2002).